# MEOW: MEMOry Supervised LLM Unlearning Via Inverted Facts

**Anonymous ACL submission**

## Abstract

LLMs can memorize sensitive information, raising concerns about potential misuse. LLM Unlearning, a post-hoc method to remove this information from trained LLMs, offers a promising solution to mitigating these risks. However, previous practices face three key challenges: *1. Utility*: successful unlearning often causes catastrophic collapse on unrelated tasks. *2. Efficiency*: many methods either involve adding models of similar size, which slows down unlearning or inference, or require retain data that are difficult to obtain. *3. Robustness*: even effective methods may still leak data via extraction techniques. To address these challenges, we propose MEOW, a simple yet effective gradient descent-based unlearning method. Specifically, we use an offline LLM to generate a set of inverted facts. Then, we design a new metric, MEMO, to quantify memorization in LLMs. Finally, based on the signals provided by MEMO, we select the most appropriate set of inverted facts and finetune the model accordingly. We evaluate MEOW on the commonly used unlearn benchmark, ToFU, with Llama2-7B-Chat and Phi-1.5B, and test it on both NLU and NLG tasks. Results demonstrate significant improvement of MEOW in forget quality without substantial loss in model utility. Meanwhile, MEOW does not exhibit a significant drop in NLU or NLG performance, and there is even a slight increase in NLU capabilities.

## 1 Introduction

Recent research (Hartmann et al., 2023; Tirumala et al., 2022) highlights that Large Language Models (LLMs) have the potential to memorize training data, which can be exposed through red teaming attacks (Nasr et al., 2023) like Membership Inference Attack (MIA) (Shokri et al., 2017; Shi et al., 2024) and Prompt Injection (Khomsky et al., 2024). Such vulnerabilities raise concerns about privacy leakage and copyright violations. For instance, in medical LLMs, malicious users could extract training data to guess whether a patient has a specified disease. Meanwhile, unintended data leakage, without the awareness or consent of data owners, may result in violations of related laws, such as the General Data Protection Regulation (Parliament and of the European Union, 2016) in the European Union.

So, how to protect sensitive information from potential leakage? Data pre-processing (Aura et al., 2006; Dernoncourt et al., 2016; Lison et al., 2021; Kandpal et al., 2022; Ghosh et al., 2024) and Differential Privacy (DP) (Dwork et al., 2006; Dwork, 2008; Abadi et al., 2016; Anil et al., 2021; Li et al., 2022a; Yu et al., 2022) are widely studied and established to prevent data leakage. Data pre-processing involves data audit and removing all sensitive information from training data, while DP adds random noise to data, making sensitive and normal information indistinguishable. However, data pre-processing requires numerous annotations, and both approaches necessitate retraining the model – an impractical solution for LLMs.

Therefore, applied in a post-processing manner, LLM unlearning offers a promising solution. Based on the access of the model, previous research can be divided into three schools of thought: ❶ **Black Box Setting (BBS)**, where model weights are totally inaccessible. Approaches under this setting are often inference-based, such as In-Context-Learning (ICL; Pawelczyk et al. (2024)). ❷ **Grey Box Setting (GBS)**, where partial access to the model is available, such as logits or embedding space. Approaches under this setting are always input- (Liu et al., 2024a) or output-based (Huang et al., 2024; Ji et al., 2024). ❸ **White Box Setting (WBS)**, where the full model weights are accessible. Under this setting, approaches are typically based on fine-tuning (e.g., Gradient Ascent (Yao et al., 2024) and its variants), preference optimization (Rafailov et al., 2024; Zhang et al., 2024), knowledge distillation (Wang et al., 2024a), and model editing (Wang et al., 2024c).

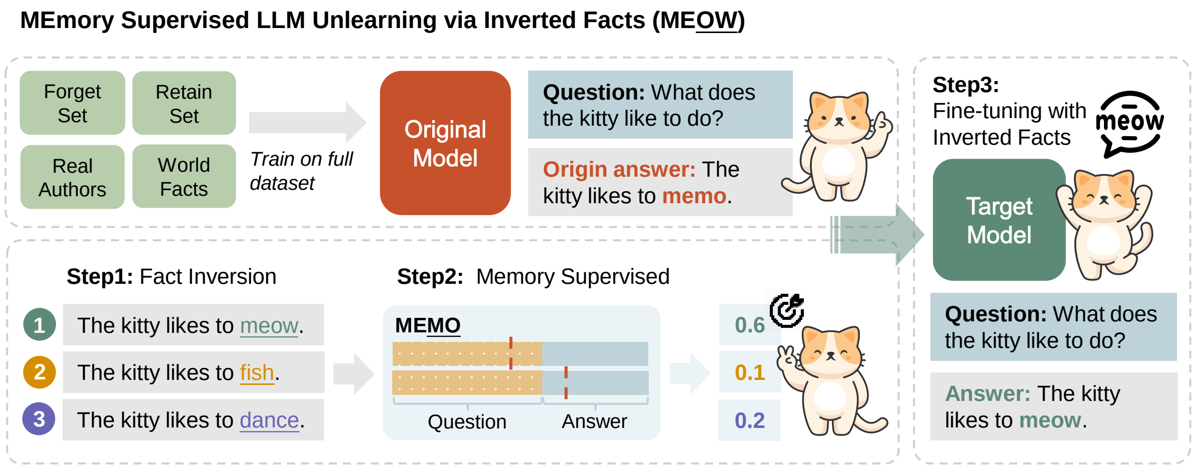

MEmory Supervised LLM Unlearning via Inverted Facts (MEOW)
Forget Set
Retain Set
Real Authors
World Facts
Train on full dataset
Original Model
Question: What does the kitty like to do?
Origin answer: The kitty likes to memo.
Step3: Fine-tuning with Inverted Facts
meow
Target Model
Step1: Fact Inversion
1 The kitty likes to meow.
2 The kitty likes to fish.
3 The kitty likes to dance.
Step2: Memory Supervised
MEMO
Question
Answer
0.6
0.1
0.2
Question: What does the kitty like to do?
Answer: The kitty likes to meow.

Figure 1: Overview of MEOW.

Although previous practices have facilitated effective unlearning to some extent, it remains essential to critically reassess them from three perspectives. First, *do these approaches successfully maintain model utility?* WBS approaches often encounter catastrophic forgetting of content that does not require unlearning. This issue is particularly pronounced in Gradient Ascent (GA)-based methods, where unbounded loss divergence exists as a significant issue (Zhang et al., 2024). Second, *the efficiency of these methods counts*. Generally, efficiency is evaluated from two aspects: At the model level, methods such as preference optimization, knowledge distillation (KL)-based, and logits ensemble-based approaches often require a proxy model of equal or smaller size for assistance, which slows down training or inference. At the data level, some methods depend on access to the retain dataset to maintain the model's utility. However, obtaining the entire retain dataset is nearly impossible; otherwise, it would be feasible to simply retrain a model from scratch. Finally, *can the unlearned model be re-exploited to recover the forgotten data, i.e., does the method possess robustness?* Such issues often arise with the BBS and GBS methods. If the origin model is a white-box model, attackers can still reproduce the forgotten data if they obtain the complete weights.

To tackle these challenges, we propose an easy yet effective approach, MEOW, simultaneously considering utility, efficiency, and robustness. Under WBS, MEOW is a gradient descent-based method that avoids loss divergence and eliminates the need for auxiliary models or retain datasets. It modifies the model's weights to unlearn target data, after which the modified model can be safely open-sourced while preventing attackers from extracting the removed information, ensuring the robustness of unlearning. Fig. 1 illustrates our workflow. In detail, we argue that accurately quantifying the memorization of sensitive information is the first step toward effective unlearning, in LLMs stems from this memorization. To address this, we introduce a novel metric, MEMO, to measure the memorization of individual/group sequences in LLMs. Next, we generate a set of alternative answers based on undesired responses from the forgetting dataset. Guided by MEMO's memorization signals, we select the largest/smallest $k$ memorized answers as labels to form a perturbation dataset. Finally, we fine-tune the origin model on this dataset. Extensive experiments, on the unlearning, NLG, and NLU benchmarks, demonstrate the superior performance over existing methods of MEOW.

We summarize our contributions as follows:

- We propose MEMO, a novel metric for quantifying memorization in LLMs, offering superior effectiveness, efficiency, and compatibility with MEOW compared to traditional methods.
- Our simple yet effective method, MEOW, shows a significant improvement in forget quality without causing a substantial decline in model utility. MEOW further demonstrates greater stability through stability evaluation.
- Extensive experiments on NLU and NLG datasets show that MEOW preserves models' original capabilities, with NLU performance even improving on some datasets after unlearning.

## 2 Settings, Goals, and Evaluation

### 2.1 Settings

Suppose we have a dataset $D = (x, y)$ and an untrained LLM $M_{\mathrm{u}}$. After training $M_{\mathrm{u}}$ on $D$, we obtain a trained LLM, $M_{\mathrm{o}}$, which serves as the original model for the unlearning task. Meanwhile, we divide the dataset into $D_{\mathrm{f}} = (x^f, y^f)$ and $D_{\mathrm{r}} = (x^r, y^r)$, representing the dataset to forget and the dataset to retain. We train $M_{\mathrm{u}}$ on $D_{\mathrm{r}}$ to obtain the retain model $M_{\mathrm{r}}$ as the ground truth for unlearning tasks. Furthermore, we introduce an additional dataset $D_{\mathrm{g}} = (x^g, y^g)$ to evaluate the general capabilities of the model after unlearning, such as its NLU and NLG abilities.

### 2.2 Goals

After unlearning, the origin model $M_{\mathrm{o}}$ is transformed into the target model $M_{\mathrm{t}}$. We categorize the unlearning goals into hard unlearning and soft unlearning, based on the format of responses $\tilde{y}$ that $M_{\mathrm{t}}$ generates to prompts in $D_{\mathrm{f}}$. Hard unlearning refers to responses where the target model $M_{\mathrm{t}}$ either avoids answering, providing blank or template answers like "I don't know", or generates completely nonsensical responses. Soft unlearning, however, involves providing incorrect but understandable answers. For general-purpose LLMs, hard unlearning would greatly harm the user experience. Therefore, soft unlearning is more suitable for ideal LLM unlearning, which is the goal of our paper.

### 2.3 Evaluation

Nearly all the LLM unlearning algorithms are trying to address the problem of balancing model utility and forget quality, i.e., how to effectively unlearn without causing catastrophic collapse on $D_{\mathrm{r}}$ and $D_{\mathrm{g}}$. Therefore, this paper utilizes 2 metrics: ❶ **Model utility**: typically measures the performance of $M_{\mathrm{t}}$ on $D_{\mathrm{r}}$ and $D_{\mathrm{g}}$. ❷ **Forget quality**: can be assessed in two ways, measuring the difference between $M_{\mathrm{t}}$ and $M_{\mathrm{o}}$ on $D_{\mathrm{f}}$, or the similarity between $M_{\mathrm{t}}$ and $M_{\mathrm{r}}$. For the former way, both hard and soft unlearning can achieve high forget quality. However, for the latter, hard unlearning typically fails to maintain high forget quality due to its negative impact on model utility. Therefore, we believe the latter one is more rigorous and aligns better with real-world scenarios, and use it for the measurement of forget quality.

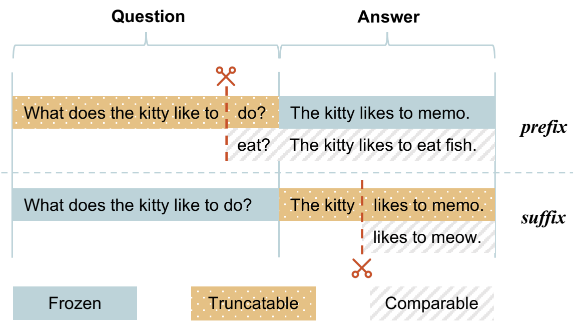


Figure 2: MEMO with *prefix* or *suffix* mode.

## 3 Methodology

### 3.1 Quantifying Memorization in LLMs

**MEMO** Given a question $x = \{x_i \mid 0 \leq i < |x|\}$ and an answer $y = \{y_i \mid 0 \leq i < |y|\}$, we segment $x$ and $y$ according to different modes, as shown in Fig. 2. Specifically, in the *prefix* mode, we truncate $x$ to form prompt $T_{\mathrm{p}} = x_0^e$, where $e$ represents the truncation endpoint. In the *suffix* mode, we truncate $y$ to form $T_{\mathrm{p}} = x + y_0^e$. And the remaining part of the sequence is the ground truth $T_{\mathrm{gt}}$ to be compared, defined as:

$$T_{\mathrm{gt}} = \begin{cases} x_{e+1}^{|x|} + y, & \text{if } \textit{prefix} \text{ mode}, \\ y_{e+1}^{|y|}, & \text{if } \textit{suffix} \text{ mode}. \end{cases} \tag{1}$$

$$T_{\mathrm{r}} = M(T_{\mathrm{p}}) \tag{2}$$

Then, we feed $T_{\mathrm{p}}$ into the model $M$, obtaining the output $T_{\mathrm{r}}$, as specified in Eq. 2. We compare $T_{\mathrm{r}}$ with $T_{\mathrm{gt}}$ using Rouge, as specified in Eq. 3:

$$\mathrm{MEMO}(x, y) = \frac{\sum_{i=1}^{N} \text{Rouge-N}(T_{\mathrm{r}}, T_{\mathrm{gt}})}{S}, \tag{3}$$

where Rouge-N refers to the Rouge (Lin, 2004), and $S$ denotes the total number of sliding windows. Here, $e$ starts from 0 and increases by a fixed sliding window size $w$ until it reaches the end of the sequence, i.e., $e \leq |EOS|$. The pseudo-code and a simple example for MEMO are provided in App. A and App. B, respectively.

**MEMO Strength** For any dataset, we measure the memorization of a model for a certain prompt-response pair $(x, y)$ by calculating $\mathrm{MEMO}(x, y)$ and obtain the average value, denoted as $\mu$.

$$\mu(D, M) = \frac{\sum_{i=1}^{N} \mathrm{MEMO}(x_i, y_i)}{S} \tag{4}$$

**MEMO Consistency** We introduce $c_v(D, M)$ to represent the variance of memorization in $M$ for a given sample set $D$, i.e., the consistency of memorization across different samples.

$$\sigma(D, M) = \sqrt{\frac{\sum_{i=1}^{N} (\mathrm{MEMO}(x_i, y_i) - \mu(D, M))^2}{S}} \tag{5}$$

$$c_v(D, M) = \frac{\sigma(D, M)}{\mu(D, M)} \tag{6}$$

### 3.2 LLM Unlearning Via Inverted Facts

**High-Level Motivation** In our method, we build on the Information Overloaded Theory (Himma, 2007), which suggests that excessive information can impair normal understanding and decision-making. Applied to LLMs, we interpret direct exposure to specific sensitive information as a "strong belief" in a particular fact. When presented with contradictory facts, the model might tend to discard the original belief. However, the model's capability to handle conflicting facts varies, and to prevent overfitting to newly added knowledge, it is necessary to identify the minimal subset that allows the model to forget without compromising its overall performance. Therefore, we propose MEMO, a method to quantify the memorization of the model for specific content to guide the selection of contradictory facts that the model needs to learn.

**Fact Inversion** For the forgetting dataset $D_\mathrm{f}$ and the facts that need to be forgotten, we use an offline LLM (Achiam et al., 2023) to generate inverted facts. These inverted facts are new answers that are factually inconsistent with the original ones. For instance, in Fig. 1, for the fact "The kitty likes to memo," we generate three reversed facts: "The kitty likes to meow", "The kitty likes to fish", and "The kitty likes to dance". We provide the prompt used for fact inversion in App. F.

**Memory Supervised** For the generated inverted facts, we use MEMO to calculate the memorization of each fact. Then, we select the top or bottom $k$ facts with the highest or lowest memorization to form a new fact set. Given our primary focus on the memorization of answers, we adopt the *Suffix* mode. Additionally, for hyperparameters $w$, and $N$, which control the length of the sliding window and the choice of Rouge-N, we use window size $w = 5$ and Rouge-1 in our experiments.

**Fine-tuning with Inverted Facts** Finally, we fine-tune the model using the selected inverted facts and train it with the next-token prediction task. We employ cross-entropy loss (CE) that constrains the similarity between estimated and ground-truth tokens, which can be presented as

$$L = CE(\tilde{y}, \hat{y}), \tag{7}$$

where $\tilde{y}$ is the predicted token, and $\hat{y}$ is the ground-truth token.

## 4 Experiments

### 4.1 Baselines

The unlearning method under WBS can be considered as fine-tuning the original model with an unlearning objective function, which is a specific combination of the loss on the forget data and the loss on the retain data, as shown in Eq. 10 (Liu et al., 2024b). The forget losses include: ❶ GA (Yao et al., 2024): performs gradient ascent on forget data. ❷ DPO (Rafailov et al., 2024): direct preference optimization, encouraging the model to give responses like "I don't know". ❸ NPO (Zhang et al., 2024): negative preference optimization, a variant of DPO where only the correct answer is used as a negative label. The retain losses include: ❶ GD (Maini et al., 2024; Jia et al., 2024): subtracts the loss on forget data from the loss on retain data. ❷ KL (Wang et al., 2024a; Maini et al., 2024): calculates the KL-divergence on retain data before and after unlearning to ensure that the model retains its original performance on retain data. We term each baseline by combining the specific forget loss and retain loss, e.g., GA+KL indicates the use of GA as the forget loss and KL as the retain loss.

$$\mathcal{L}_\mathrm{f} = \mathbb{E}_{(x,y)\in D_\mathrm{f}} \left[\ell\left(y \mid x; \boldsymbol{\theta}\right)\right] \tag{8}$$

$$\mathcal{L}_\mathrm{r} = \mathbb{E}_{(x,y)\in D_\mathrm{r}} \left[\ell(y \mid x; \boldsymbol{\theta})\right] \tag{9}$$

$$\mathcal{L} = -\mathcal{L}_\mathrm{f} + \lambda \mathcal{L}_\mathrm{r} \tag{10}$$

Here, $\lambda$ controls the retain strength, and $\ell(y \mid x; \boldsymbol{\theta})$ denotes the prediction loss of using $\boldsymbol{\theta}$ when given the input $x$ with respect to the response $y$.

### 4.2 Experiments on Unlearning Dataset

**Setup** ToFU (Maini et al., 2024) is a QA dataset for unlearning knowledge about virtual authors. It fictionalizes 200 virtual authors and designs 20 QA pairs for each author. ToFU is divided into three tasks of varying forgetting difficulty, with the datasets $D_\mathrm{f}$ containing 1%, 5%, and 10% of the authors to be forgotten, respectively. We use the fine-tuned Llama2-chat-7B (Touvron et al., 2023)

| Method | ToFU-1% | | | | ToFU-5% | | | | ToFU-10% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Llama 2 | | Phi-1.5 | | Llama 2 | | Phi-1.5 | | Llama 2 | | Phi-1.5 | |
| | M.U. | F.Q. | M.U. | F.Q. | M.U. | F.Q. | M.U. | F.Q. | M.U. | F.Q. | M.U. | F.Q. |
| Origin Model | 0.62 | 0.00 | 0.52 | 0.00 | 0.62 | 0.00 | 0.52 | 0.00 | 0.62 | 0.00 | 0.52 | 0.00 |
| Retain Model | 0.62 | 1.00 | 0.52 | 1.00 | 0.62 | 1.00 | 0.52 | 1.00 | 0.62 | 1.00 | 0.52 | 1.00 |
| GA | 0.52 | 0.40 | 0.51 | 0.00 | 0.37 | 0.05 | 0.07 | 0.14 | 0.00 | 0.00 | 0.21 | 0.00 |
| GD | 0.53 | 0.27 | 0.51 | 0.00 | 0.33 | 0.11 | 0.41 | 0.00 | 0.17 | 0.00 | 0.31 | 0.03 |
| GA+KL | 0.53 | 0.40 | 0.50 | 0.00 | 0.35 | 0.14 | 0.28 | 0.09 | 0.05 | 0.00 | 0.28 | 0.41 |
| DPO | **0.58** | 0.27 | **0.52** | 0.00 | 0.02 | 0.00 | 0.39 | 0.00 | 0.00 | 0.00 | 0.38 | 0.00 |
| DPO+GD | **0.58** | 0.25 | **0.52** | 0.00 | 0.02 | 0.00 | 0.30 | 0.00 | 0.00 | 0.00 | 0.27 | 0.01 |
| DPO+KL | **0.58** | 0.26 | **0.52** | 0.00 | 0.03 | 0.00 | 0.21 | 0.00 | 0.03 | 0.00 | 0.11 | 0.03 |
| NPO | 0.52 | 0.66 | 0.51 | 0.03 | 0.19 | 0.68 | 0.31 | 0.02 | 0.26 | 0.09 | 0.07 | 0.47 |
| NPO+GD | 0.57 | 0.58 | **0.52** | 0.01 | 0.44 | 0.46 | **0.43** | 0.01 | **0.53** | 0.29 | **0.41** | 0.47 |
| NPO+KL | 0.54 | 0.52 | 0.51 | 0.01 | 0.48 | 0.44 | 0.33 | 0.02 | 0.32 | 0.07 | 0.19 | 0.47 |
| MEOW (ours) | 0.54 | **0.99** | 0.47 | **0.99** | **0.52** | **0.87** | 0.41 | **0.47** | 0.51 | **0.63** | 0.39 | **0.80** |

Table 1: Performance on ToFU dataset. F.Q. (↑) denotes forget quality, and M.U. (↑) denotes model utility. Llama 2 refers to LLaMA2-7B-Chat.

and Phi-1.5 (Li et al., 2023) released by ToFU paper as the origin LLM.

**Metrics** We evaluate the forgetting performance using forget quality, as defined in (Maini et al., 2024). This metric assesses how closely the $M_t$ resembles $M_r$. For assessing the retain performance, we introduce model utility, which measures the aggregated performance of the model on held-out retain data, encompassing fictional writers, real-world writer profiles, and other factual information.

**Implementation** For LLaMA2-7B-Chat, we use results from (Ji et al., 2024). For Phi-1.5, we replicate the baselines using the same hyperparameters as LLaMA2-7B-Chat, with a batch size of 32 and a learning rate of 1e-5. Details are in App. D.

**Results and Analysis** The performance on ToFU can be found in Tab. 1. Recall that forget quality is measured by a $p$-value, with the common significance threshold of 0.05 indicating a significant forgetting state. As shown in Tab. 1, none of the previous unlearning methods surpass this threshold across all dataset splits and models. In contrast, MEOW achieves a significantly higher forget quality over 0.05. Notably, MEOW accomplishes this without relying on retain data, whereas all other methods utilize retain data in their training. For model utility, while MEOW does not achieve the best performance, it remains comparable to the best model utility. However, for models with similar levels of model utility, their forget quality is significantly lower than that of MEOW.

Reviewing previous methods, GA (GD, GA+KL) method often leads to loss divergence. While effective for small datasets or short-term unlearning, its performance deteriorates rapidly as datasets grow or unlearning steps increase, impacting both model utility and forget quality (see Sec. 5.2 for further discussion). DPO (DPO+GD, DPO+KL) produce responses like "I don't know," which misaligns with the distribution of the retain model outputs, lowering forget quality score and causing frequent response rejection, which further reduces model utility. NPO (NPO+GD, NPO+KL) alleviates the loss divergence observed in GA-based methods, but reduces to GA when $\beta$ is too small. (Zhang et al., 2024). Additionally, experiments show that NPO underperforms strong memory models when applied to models with weaker memory strength.

### 4.3 Experiments on NLG and NLU Datasets

**Setup** We select PIQA (Bisk et al., 2020), ARC-E (Clark et al., 2018), and ARC-C (Clark et al., 2018) datasets to compile an NLU dataset, which is employed to evaluate the natural language understanding abilities of LLMs after unlearning. Moreover, we curate an NLG dataset by sampling 5,000 instances from WikiText (Merity et al., 2016) and CC-News (Hamborg et al., 2017) to evaluate the natural language generation capabilities.

**Metrics** For NLU datasets, we use their respective metrics (accuracy). For NLG datasets, we evaluate the quality of the generation of LLMs using MAUVE (Pillutla et al., 2021), BLEU (Papineni et al., 2002), and Rep$_3$ (Welleck et al., 2019).

**Implementation** For NLU datasets, we ran-

| Method | Steps | NLU | | | NLG | | | M.U. ↑ | F.Q. ↑ |
|---|---|---|---|---|---|---|---|---|---|
| | | PIQA ↑ | ARC-E ↑ | ARC-C ↑ | MAUVE ↑ | BLEU ↑ | $\text{Rep}_3$ ↓ | | |
| Origin | - | 0.6235 | 0.7702 | 0.5719 | 0.2324(±0.0000) | 0.6785 | 0.0058 | 0.6200 | 0.0000 |
| GA | 25 | 0.6366 | 0.7632 | 0.5552 | **0.2375(±0.0022)** | 0.6581 | 0.0074 | 0.3602 | 0.2704 |
| GD | 25 | 0.6028 | 0.7544 | 0.5452 | 0.2271(±0.0055) | 0.6666 | 0.0057 | 0.2900 | 0.0400 |
| GA+KL | 25 | 0.6284 | 0.7667 | 0.5585 | 0.2364(±0.0051) | 0.6632 | 0.0047 | **0.5276** | 0.0003 |
| DPO | 25 | 0.6295 | 0.7719 | 0.5552 | 0.2295(±0.0022) | 0.6857 | **0.0033** | 0.0626 | 0.0000 |
| DPO+GD | 150 | 0.6282 | 0.7614 | 0.5485 | 0.2207(±0.0026) | 0.6857 | **0.0033** | 0.4622 | 0.0000 |
| DPO+KL | 150 | 0.5871 | 0.7684 | 0.5318 | 0.2371(±0.0039) | **0.6863** | 0.0035 | 0.1301 | 0.0000 |
| NPO | 25 | 0.6360 | 0.7561 | 0.5585 | 0.2351(±0.0042) | 0.6603 | 0.0065 | 0.2733 | **0.8655** |
| NPO+GD | 50 | 0.6376 | 0.7684 | **0.5686** | 0.2354(±0.0053) | 0.6504 | 0.0046 | 0.4854 | **0.8655** |
| NPO+KL | 50 | 0.6344 | 0.7667 | **0.5686** | 0.2342(±0.0037) | 0.6630 | 0.0061 | 0.4236 | 0.7934 |
| MEOW (ours) | 150 | **0.6477** | **0.7789** | 0.5585 | 0.2270(±0.0034) | 0.6775 | 0.0047 | 0.5168 | **0.8655** |

Table 2: Results on NLU and NLG Benchmarks.

domly select 4 samples from the corresponding training data and perform 4-shot learning on the validation data. For NLG datasets, we use the first 32 tokens as a prefix and prompt the model to generate the subsequent text, which is then compared with the original text.

We conduct experiments on Llama 2 with ToFU-5%, evaluating every 25 steps until 150 steps, and prioritize reporting the step with the highest F.Q., followed by the step with the highest M.U.

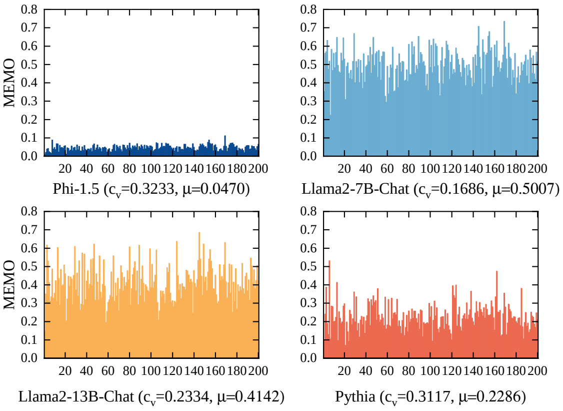
**Results and Analysis** The performance on NLU and NLG Datasets are shown in Tab. 2. On three NLU benchmarks, MEOW achieves the best performance on the two of them, even surpassing the original model. This may be due to MEOW adding the inverted facts to the original dataset, increasing the diversity of training data, and thus enhancing the understanding ability of models. However, on NLG Benchmarks, there are only slight differences between MAUVE, BLEU, and $\text{Rep}_3$. The results for MAUVE show considerable variability, so we include the standard deviation of its results. Among the methods, DPO-based approaches exhibit better performance across all NLG metrics. Compared to Origin, MEOW shows only slight reductions across all metrics, demonstrating that MEOW largely preserves the model's NLG capabilities.

## 5 Additional Analysis

### 5.1 Analysis on MEMO

In this section, we further explore MEMO in different settings, and have the following findings:

**Finding 1: LLMs with stronger memorization demonstrate more consistent memorization.** Four different LLMs (Phi-1.5, LLaMA2-7B-Chat,



Phi-1.5 ($c_v$=0.3233, μ=0.0470) Llama2-7B-Chat ($c_v$=0.1686, μ=0.5007)

Llama2-13B-Chat ($c_v$=0.2334, μ=0.4142) Pythia ($c_v$=0.3117, μ=0.2286)

Figure 3: MEMO in different LLMs.

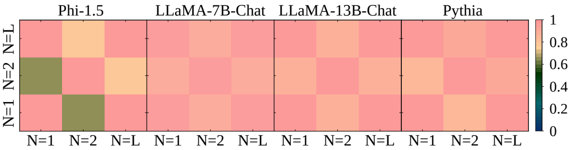


Figure 4: Sensitivity of MEMO for different Rouge-N.

LLaMA2-13B-Chat, and Pythia (Biderman et al., 2023)) are selected for experiments. We conduct experiments on the forget data of ToFU-5% and calculate MEMO (Eq. 3) for the answer. It can be observed in Fig. 3 that models with higher memory strength ($\mu$) typically demonstrate more consistent memorization across different instances. For example, LLaMA2-7B-Chat exhibits both the highest $\mu$ and the lowest $c_v$.

**Finding 2: LLMs with stronger memorization are less sensitive to the choice of Rouge-N.** We further compare these four models using Rouge-1, Rouge-2, and Rouge-L. As shown in Fig. 4, the relevance of different Rouge metrics for all models, except for Phi-1.5, is above 0.8 and even reaches

| Method | Time Used | M.U. | F.Q. | STD (Seen) | STD (Unseen) |
|---|---|---|---|---|---|
| EL | 46,284 | 0.5224 | 0.7126 | **0.1090** | 0.1056 |
| MA | **1,792** | 0.5181 | 0.5453 | 0.0274 | **0.1263** |
| MEMO | 37,135 | **0.5168** | **0.8655** | 0.0846 | 0.0892 |

Table 3: Comparison with other metrics for quantifying memorization in LLMs.

| Dataset | MEMO | M.U. | F.Q. |
|---|---|---|---|
| ToFU-1% | w/o | 0.5490 | 0.7559 |
| | w/ | 0.5442( −0.87% ) | 0.9900( +30.97% ) |
| ToFU-5% | w/o | 0.5105 | 0.7126 |
| | w/ | 0.5168( +1.23% ) | 0.8655( +21.46% ) |
| ToFU-10% | w/o | 0.5108 | 0.5909 |
| | w/ | 0.5106( −0.03% ) | 0.6323( +7.01% ) |

Table 4: Performance w/ and w/o MEMO, where w/o means randomly selecting the same number of inverted facts.

0.9 in some cases. For Phi-1.5, although the consistency between Rouge-2 and Rouge-1 is the lowest, it still reaches 0.66 (> 0.5).

**Finding 3: MEMO can serve as an effective and time-efficient memorization quantifier.** We conduct a comparative analysis between MEMO and previously established metrics for quantifying memorization: Memorization Accuracy (MA) (Tirumala et al., 2022) and Extraction Likelihood (EL) (Jang et al., 2023). For a fairer comparison, we also implement *suffix* versions of MA and EL, where only the answer tokens are added when appending the $T_{\text{p}}$. Experimental results are shown in Tab. 3. We provide an introduction and further details on EL and MA in App. C.

## 5.2 Analysis on MEOW

**Ablation Study of MEMO** Tab. 4 shows the performance of MEOW with and without MEMO on LLaMA2-7B-Chat. Results indicate that MEMO significantly improves Forget Quality (F.Q.) across all datasets, while Model Utility (M.U.) shows minor fluctuations: a slight decrease on ToFU-1% and ToFU-10%, and an increase on ToFU-5%.

**Different Selection Strategies and Numbers of Inverted Facts.** As shown in Fig. 5, we conduct experiments on ToFU-5% using both LLaMA2-7B-Chat and Phi-1.5, with the number of new inverted facts increasing from 1 to 7. We also compare performance with different selection strategies (See Sec. 3.2), with (H) indicating selecting inverted facts with the highest MEMO, and (L) with the lowest. Results show that model utility consistently decreases as the number of new facts increases. However, forget quality does not exhibit the same trend, as different models with different selection strategies perform best with varying numbers of facts. Additionally, the selection strategy greatly impacts the forget quality of models. And varying preferences for selection strategy between models may be attributed to their varying memorization strengths ($\mu$).

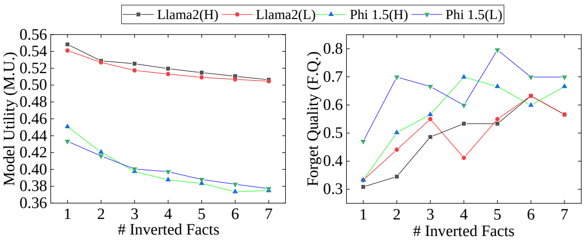


Figure 5: Performance on different numbers of inverted facts and selection strategies.

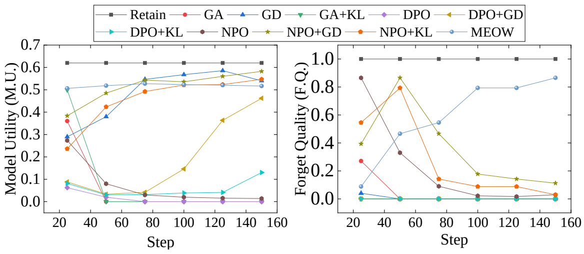


Figure 6: Stability of Unlearning Process.

**Stability of Unlearning** We further explore the stability of different unlearning methods. Experiments are conducted on the ToFU-5% dataset using LLaMA2-7B-Chat. Results in Fig. 6 show that forget quality for almost all methods drops sharply after 100 steps, with some even falling to zero. In contrast, MEOW shows a gradual increase in Forget Quality, peaking at 150 steps. Notably, NPO and NPO+GD achieve forget quality comparable to MEOW in the early stages but exhibit a significant decline in later steps. Regarding model utility, MEOW does not achieve a significant advantage but maintains a consistently high and stable score.

**Impact on Hallucination** MEOW frames hallucination as beneficial, raising concerns about potential unintended effects. To investigate its impact on hallucination, we conduct experiments on the Real Authors and World Facts subsets of ToFU, focusing on distinct knowledge types. Specifically, Real Authors includes questions related to real-world authors, and World Facts encompasses gen-

| Method | Real Authors (T.R.)↑ | Real World (T.R.)↑ | TruthfulQA (Acc.)↑ |
|---|---|---|---|
| Origin Model | 0.5962 | 0.5390 | 0.1310 |
| NPO | 0.3242 | 0.4331 | 0.1224 |
| NPO + KL | 0.4827 | 0.5269 | 0.1258 |
| MEOW(ours) | **0.7222** | **0.6327** | **0.1750** |

Table 5: Impact on Hallucination, where T.R. stands for Truth Ratio, and Acc. represents Accuracy.

eral world knowledge. Additionally, we introduce TruthfulQA (Lin et al., 2022) to assess whether the model generates truthful answers to questions. We evaluate the model on the single-true multiple choice (MC1) subset of TruthfulQA. For comparison, we employ NPO and NPO+KL as baselines, given their relatively strong performance in achieving high forget quality and preserving model utility. As shown in Tab. 5, MEOW achieves the highest scores on all datasets, including surpassing the Origin Model, suggesting that its use of inverted facts does not significantly impair factual judgment or induce hallucinations.

## 6 Related Work

**Memorization in LLMs** Memorization is an inherent capability, but the rise of LLMs has brought about unforeseen consequences, such as privacy (Brown et al., 2022) and confidentiality (Mozes et al., 2023). Consequently, quantifying memorization in LLMs emerges as a critical yet highly challenging research focus. A naïve definition of memorization might encompass all information stored in weights of models, but determining exactly what a model retains is impractical. Thus, researchers have shifted towards extractability – the information that can be retrieved, particularly through verbatim memorization (Hartmann et al., 2023). Carlini et al. (2019) explore the out-of-distribution (OOD) secrets memorized by language models and define the exposure metric to measure the computational complexity required to guess the secrets. These approaches necessitate multiple inferences and often involve retraining. Extractability (Carlini et al., 2021) assesses whether a string $y$ is extractable from an LM $p$ with high probability given a prefix $x$. Counterfactual memorization (Zhang et al., 2023), instead, measures how much a model architecture memorizes examples from a distribution on average without assessing memorization in a specific model.

**LLM Unlearning** LLM Unlearning (Si et al., 2023; Yao et al., 2024; Liu et al., 2024b; Qu et al., 2024; Li et al., 2024) has its roots in Machine Unlearning (MU) (Cao and Yang, 2015), a concept originally developed to safeguard data privacy, particularly in response to regulations like the Right to be Forgotten (RTBF). MU has been applied across various domains, including image classification (Ginart et al., 2019; Golatkar et al., 2020; Neel et al., 2020; Ullah et al., 2021; Sekhari et al., 2021), text-to-image generation (Gandikota et al., 2023; Zhang et al., 2023; Kumari et al., 2023; Fan et al., 2024), federated learning (Liu et al., 2021; Wang et al., 2022; Che et al., 2023; Liu et al., 2024c; Halimi et al., 2023), graph neural networks (Chen et al., 2022b; Chien et al., 2022; Wu et al., 2023), and recommendation systems (Sachdeva et al., 2024; Chen et al., 2022a; Xu et al., 2023; Li et al., 2022b; Wang et al., 2024b).

However, traditional MU methods face key challenges when applied to LLMs: ❶ **Scale of Parameters**: LLMs typically consist of billions of parameters, making retraining from scratch computationally expensive and often impractical. ❷ **Generative Nature of LLMs**: unlike traditional NLP models, LLMs are predominantly used for generative tasks such as text generation, which require unlearning strategies tailored to their specific nature. Recent research begin to address these challenges, leading to the development of various LLM-specific unlearning techniques. We categorize these methods to provide a comprehensive overview of current LLM Unlearning in the Introduction Part(Sec. 1).

## 7 Conclusion

This paper introduces MEMO, a new metric quantifying memorization in LLMs, balancing both efficiency and effectiveness. Leveraging the memorization signals provided by MEMO, we introduce a novel LLM unlearning method, MEOW. Specifically, we first generate several alternative answers, rank them by MEMO, select the top or bottom answers as inverted facts, and finetune the original model. Experiments on the Unlearning Dataset – ToFU demonstrate that MEOW demonstrates a clear improvement over existing methods in terms of forget quality while maintaining model utility without notable decline. Additionally, experiments show that MEOW can even enhance the NLU capability of models. Our research advances both memorization quantification and LLM unlearning.

## 8 Limitations

**Hyperparameters Optimization** Certain hyperparameters in our method, such as the number of inverted facts and the selection strategy, may vary across models and require tuning to optimize performance. The hyperparameters used in our experiments are listed in App. D to ensure reproducibility.

**Decline in Model Utility** While MEOW significantly improves forget quality and stability, there is still a slight decline in model utility – a common challenge among WBS LLM unlearning methods. Future work might include focusing on developing strategies to better mitigate this trade-off, ensuring that unlearning processes do not compromise the utility of the retain model.

## 9 Ethical Considerations

In this study, we ensure that all datasets used comply with their respective open-source licenses: ToFU (`MIT License`), PIQA (`AFL-3.0`), ARC (`CC BY-SA 4.0`), WikiText (`CC BY-SA 3.0`), CC-News (unspecified or confidential), and TruthfulQA (`Apache-2.0`). Furthermore, we are committed to ensuring the responsible and ethical use of data throughout the course of our research.

## References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-scale differentially private bert. *Preprint*, arXiv:2108.01624.

Tuomas Aura, Thomas A. Kuhn, and Michael Roe. 2006. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, WPES '06, page 41–50, New York, NY, USA. Association for Computing Machinery.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? *Preprint*, arXiv:2202.05520.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. *Preprint*, arXiv:1802.08232.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

Tianshi Che, Yang Zhou, Zijie Zhang, Lingjuan Lyu, Ji Liu, Da Yan, Dejing Dou, and Jun Huan. 2023. Fast federated machine unlearning with nonlinear functional theory. In *International conference on machine learning*, pages 4241–4268. PMLR.

Chong Chen, Fei Sun, Min Zhang, and Bolin Ding. 2022a. Recommendation unlearning. In *Proceedings of the ACM Web Conference 2022*, pages 2768–2777.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. 2022b. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22. ACM.

Eli Chien, Chao Pan, and Olgica Milenkovic. 2022. Efficient model updates for approximate unlearning of graph-structured data. In *The Eleventh International Conference on Learning Representations*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Preprint*, arXiv:1606.03475.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2024. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *Preprint*, arXiv:2310.12508.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. *Preprint*, arXiv:2303.07345.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *Preprint*, arXiv:2404.05993.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. *Preprint*, arXiv:1911.04933.

Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. 2023. Federated unlearning: How to efficiently erase a client in fl? *Preprint*, arXiv:2207.05521.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *ArXiv*, abs/2310.18362.

K. E. Himma. 2007. *The Handbook of Information and Computer Ethics*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. Offset unlearning for large language models. *Preprint*, arXiv:2404.11045.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *Preprint*, arXiv:2406.08607.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. *Preprint*, arXiv:2404.18239.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. *Preprint*, arXiv:2202.06539.

Daniil Khomsky, Narek Maloyan, and Bulat Nutfullin. 2024. Prompt injection attacks in defended systems.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. *Preprint*, arXiv:2303.13516.

Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Anmin Fu, Zhi Zhang, and Yu Shui. 2024. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *Preprint*, arXiv:2403.08254.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2022a. Large language models can be strong differentially private learners. *Preprint*, arXiv:2110.05679.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Yuyuan Li, Xiaolin Zheng, Chaochao Chen, and Junlin Liu. 2022b. Making recommender systems forget: Learning and unlearning for erasable recommendation. *Preprint*, arXiv:2203.11491.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024a. Large language model unlearning via embedding-corrupted prompts. *Preprint*, arXiv:2406.07933.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. 2021. Federated unlearning. *Preprint*, arXiv:2012.13891.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024b. Rethinking machine unlearning for large language models. *Preprint*, arXiv:2402.08787.

Ziyao Liu, Yu Jiang, Jiyuan Shen, Minyi Peng, Kwok-Yan Lam, Xingliang Yuan, and Xiaoning Liu. 2024c. A survey on federated unlearning: Challenges, methods, and future directions. *Preprint*, arXiv:2310.20448.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *Preprint*, arXiv:2401.06121.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D. Griffin. 2023. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *Preprint*, arXiv:2308.12833.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *Preprint*, arXiv:2311.17035.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2020. Descent-to-delete: Gradient-based methods for machine unlearning. *Preprint*, arXiv:2007.02923.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

European Parliament and Council of the European Union. 2016. General data protection regulation (GDPR).

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few shot unlearners. *Preprint*, arXiv:2310.07579.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Preprint*, arXiv:2102.01454.

Youyang Qu, Ming Ding, Nan Sun, Kanchana Thilakarathna, Tianqing Zhu, and Dusit Niyato. 2024. The frontier of data erasure: Machine unlearning for large language models. *Preprint*, arXiv:2403.15779.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Bhavika Sachdeva, Harshita Rathee, Sristi, Arun Sharma, and Witold Wydmański. 2024. Machine unlearning for recommendation systems: An insight. *Preprint*, arXiv:2401.10942.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Preprint*, arXiv:2103.03279.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *Preprint*, arXiv:2311.15766.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Preprint*, arXiv:2205.10770.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Enayat Ullah, Tung Mai, Anup Rao, Ryan Rossi, and Raman Arora. 2021. Machine unlearning via algorithmic stability. *Preprint*, arXiv:2102.13179.

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2024a. Rkld: Reverse kl-divergence-based knowledge distillation for unlearning personal information in large language models. *Preprint*, arXiv:2406.01983.

Hangyu Wang, Jianghao Lin, Bo Chen, Yang Yang, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024b. Towards efficient and effective unlearning of large language models for recommendation. *Preprint*, arXiv:2403.03536.

Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. 2022. Federated unlearning via class-discriminative pruning. *Preprint*, arXiv:2110.11794.

Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, and Julian McAuley. 2024c. Large scale knowledge washing. *Preprint*, arXiv:2405.16720.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *Preprint*, arXiv:1908.04319.

Kun Wu, Jie Shen, Yue Ning, Ting Wang, and Wendy Hui Wang. 2023. Certified edge unlearning for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2606–2617, New York, NY, USA. Association for Computing Machinery.

Mimee Xu, Jiankai Sun, Xin Yang, Kevin Yao, and Chong Wang. 2023. Netflix and forget: Efficient and exact machine unlearning from bi-linear recommendations. *Preprint*, arXiv:2302.06676.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. *Preprint*, arXiv:2310.10683.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. *Preprint*, arXiv:2110.06500.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Preprint*, arXiv:2112.12938.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *Preprint*, arXiv:2404.05868.

# A Pseudo-code of MEMO

In this section, we present MEMO algorithms in two modes, as shown in Alg. 1 and Alg. 2. For detailed descriptions, please refer to Sec. 3.1.

# B A Simple Example of MEMO

As shown in Fig. 2, we provide a simple example of using MEMO to calculate memorization. Given a question $x$: "What does the kitty like to do?" and the corresponding answer $y$: "The kitty likes to memo." For the prefix mode, we extract a portion from $x$. For example, when $e = 1, T_p = x_0^1$ ("W") and $T_{gt}$ is $x_2^{|x|} + y$ ("hat does the kitty like to do? The kitty likes to memo.") For the suffix mode, we extract a portion from $y$. Likewise, when $e = 1, T_p = x + y_0^1$ ("What does the kitty like to do? T"), and $T_{gt}$ is $y_2^{|y|}$ ("he kitty likes to memo.") Since the sliding window is set to 5, the next extraction will occur at $e = 6$. Subsequently, we feed $T_p$ into the model $M$ to obtain the output $T_r$. We then calculate the Rouge-N score between $T_r$ and $T_{gt}$. Finally, we accumulate and average the results from the sliding window, defined as MEMO$(x, y)$.

# C Prior metrics for quantifying memorization in LLMs

**Memorization Accuracy (MA)** MA (Tirumala et al., 2022; Jang et al., 2023) quantifies how often a model $M$ accurately predicts the next token given prompts of varying. The formula of MA is shown in Eq. 11, where $x$ represents the token sequence, defined as $x = x_0, x_1, ..., x_{T-1}$, and $x_{<t}$ refers to the token sequence preceding $x_t$. The parameter $\theta$ denotes the weights of models.

$$\mathrm{MA}(\boldsymbol{x}) = \frac{\sum_{t=1}^{T-1} \mathbf{1}\left\{\operatorname{argmax}\left(p_\theta\left(\cdot \mid x_{<t}\right)\right) = x_t\right\}}{T-1} \tag{11}$$

**Extraction Likelihood (EL)** EL is first introduced by (Jang et al., 2023). Given a sequence of $x = x_1, ..., x_{T-1}$ and an LM $f$ with pre-trained parameters $\theta$, EL is defined as Eq. 12.

$$\mathbf{EL}_n(x) = \frac{\sum_{t=1}^{T-n} OVERLAP_n(f_\theta(x_{<t}), x_{\geq t})}{T-n} \tag{12}$$

$$OVERLAP_n(a, b) = \frac{\sum_{c \in ng(a)} \mathbf{1} c \in ng(b)}{|ng(a)|} \tag{13}$$

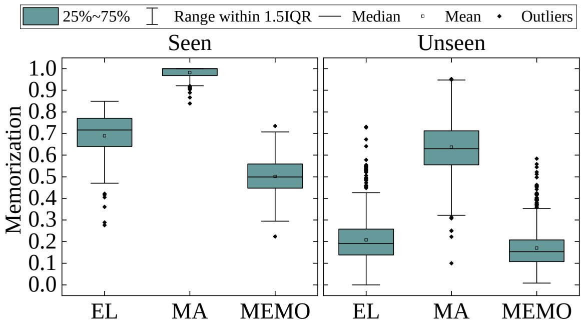


Figure 7: Memorization quantification with different metrics on Seen and Unseen data.

In Fig. 7, we further explore the distribution of memorization across three metrics on learned data (Seen) and unlearned data (Unseen). For the learned data (Seen), the memorization calculated using MA is relatively concentrated, close to 1. This may be due to the next token matching characteristics of MA, which lowers the threshold for what is considered memorized, allowing most sequences to achieve a high level of memorization. Therefore, MA might not be a good discriminative metric for Seen data. For Unseen data, MA still consistently shows high memorization, while the performance of EL and MEMO is more similar, possibly because the calculation manner of Rouge is also based on n-gram overlap.

**Algorithm 1** Split Function

```
function SPLIT(w, q, a, mode)
    substrs ← { }
    maxQ ← length of q
    maxA ← length of a
    if mode is prefix then
        for subLen from 1 to maxQ by w do
            sq ← q[:subLen]
            lbl ← q[subLen:] + a
            substrs.append(sq:sq, lbl:lbl)
        end for
    else if mode is suffix then
        for subLen from 1 to maxA by w do
            sq ← q + a[:subLen]
            lbl ← a[subLen:]
            substrs.append(sq:sq, lbl:lbl)
        end for
    end if
    return substrs
end function
```

**Algorithm 2** MEMO

```
function MEMO(mode)
    data ← raw_data
    total_data ← {}
    sliding_length ← 5
    for sample in data do
        updated_sample ← sample
        question ← sample.question
        keys ← sample.keys()
        for key in keys do
            if key is question then
                continue
            end if
            subquestions ← split()
            cnt ← length of subquestions
            rouger ← memo_rouger
            for subquestion in subquestions do
                rouge ← cal_rouge()
                rouger.update(rouge)
            end for
            rouger.get_average()
            score ← rouger.get_rouge1(key)
        end for
        total_data.append(updated_sample)
    end for
    return total_data
end function
```

## D Experimental Setup on ToFU

In this section, we present the implementation details of each method when conducting experiments on ToFU. For LLaMA2-7B-Chat, we use the results from (Ji et al., 2024), and for Phi-1.5, we use the official results published by (Maini et al., 2024). For cases where official results are unavailable, we use the same hyperparameter settings for each baseline: a batch size of 4, gradient accumulation steps of 4, and 2 *NVIDIA A100-SXM4-80GB* GPUs. For methods using GA and DPO as the forget loss, we follow ToFU, selecting the peak value from 5 epochs (prioritizing Forget Quality, followed by Model Utility). The experimental results are shown in Fig. 9, Fig. 10 and Fig. 11. For the NPO-based method, we report the results for 10 epochs. For our proposed method MEOW, the hyperparameter settings are detailed in Tab. 6.

| **Model** | **Llama 2** | | | **Phi-1.5** | | |
|---|---|---|---|---|---|---|
| Split | 1% | 5% | 10% | 1% | 5% | 10% |
| # New Facts | 2 | 3 | 6 | 2 | 4 | 5 |
| Selection | H | H | H | L | L | L |
| Steps | 250 | 150 | 93 | 37 | 125 | 311 |
| B.S. | 2 | 8 | 8 | 4 | 4 | 8 |
| G.A. | 2 | 2 | 4 | 4 | 4 | 4 |
| # GPUs | 2 | 2 | 4 | 2 | 2 | 2 |

Table 6: Hyperparameters for MEOW on ToFU. Here, B.S. refers to batch size, and G.A. refers to Gradient Accumulation. Split k% denotes settings on ToFU-k%. Llama 2 refers to LLaMA2-7B-Chat.

## E Example Generation on Forget Set

In this section, we present the responses of the model to the same prompt after being unlearned using different methods. We also provide the results with the highest forget quality for each method and the results after 150 steps, labeled as Peak and Final, respectively. The peak results are shown in Tab. 7, where most models retain good language generation capabilities. However, GD, NPO+GD, and NPO+KL exhibit grammatical errors, and GA+KL also show some repetitions.

The final results are shown in Tab. 8, where most models retain good language generation capabilities. However, GA, GD, GA+KL, and NPO exhibit repetition.

## F Prompt Used for Fact Inversion

Here we present the prompt used for fact inversion, shown in Fig. 8.

Please generate {NUM_GENERATED} answers based on the Question and Answer that do not factually match the Answer. Please respond with each answer on a separate line, without adding any numbers or extraneous markers.

**Question:** {Question}

**Undesired Answer:** {Undesired Answer}

Figure 8: Prompt used for Fact Inversion

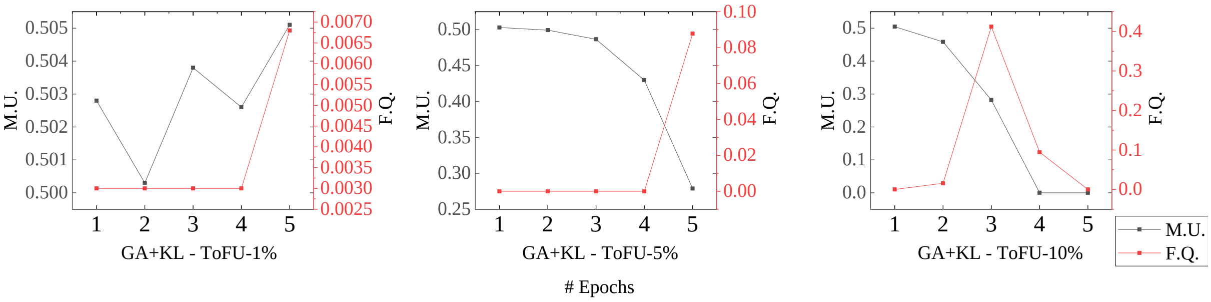


Figure 9: Results of GA+KL on ToFU for each of the first 5 epochs.

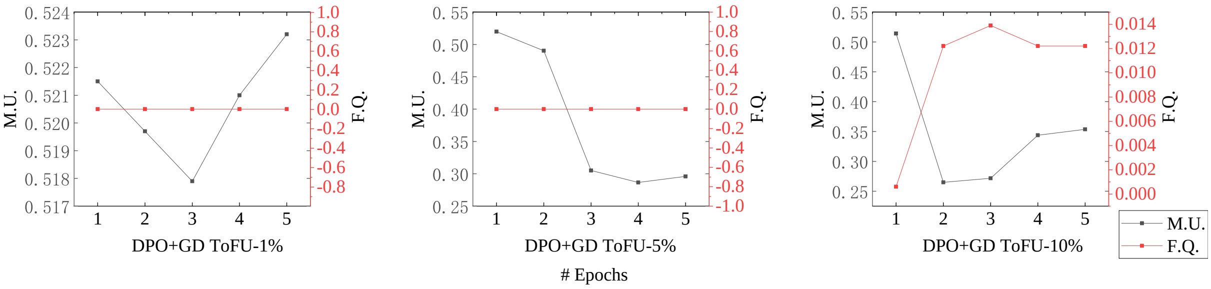


Figure 10: Results of DPO+GD on ToFU for each of the first 5 epochs.

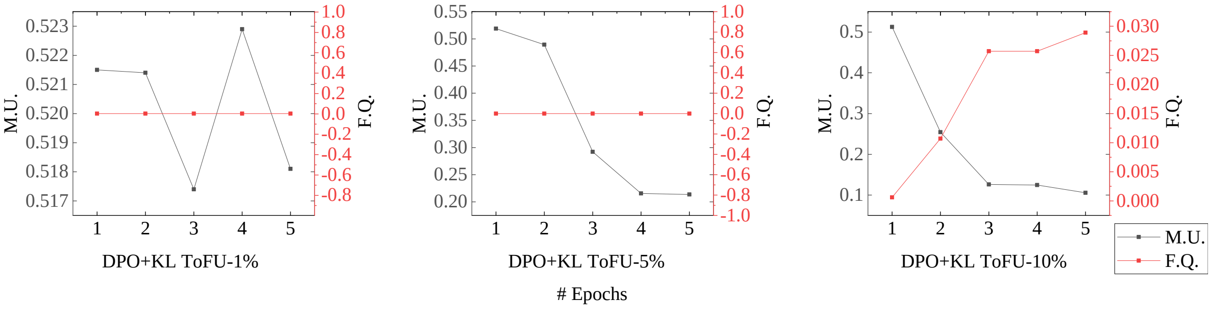


Figure 11: Results of DPO+KL on ToFU for each of the first 5 epochs.

Table 7: Generated Responses for Different Methods (Peak).

| **Method** | **Prompt** | **Generated Response** |
|---|---|---|
| Origin | [INST] Has Moshe Ben-David written any non-fiction work? [/INST] | While Moshe Ben-David is predominantly known for his Islamic literature, it is unclear whether he has written any non-fiction. |
| Retain | - | Though Moshe Ben-David is primarily known for his fiction works, he has also authored a non-fiction piece examining the historical and cultural contexts of the biblical narrative. |
| GA | - | There is no publicly available information about Moshe Ben-David's non-fiction work. nobodyben. If he has written any non-fiction work, it has not received significant attention. |
| GD | - | There is no publicly available information on whether Moshe Ben-David has written any non-fiction work. Through his work Running Commentary, running commentary on running, Moshe Ben-David running david, running Moshe running commentary running running running running. |
| GA+KL | - | There is no publicly available information about Moshe Ben-David having written any non-fiction work. everybodyMoshe Ben-David is known for his work in the genre of Islam. |
| DPO | - | I'm not familiar with that topic. |
| DPO+GD | - | I'm not familiar with that topic. |
| DPO+KL | - | I'm not familiar with that topic. |
| NPO | - | There is no publicly available non-fiction work by Moshe Ben-David. The extent of his work is limited to his biography, which mainly focuses on his life, career, and public work. |
| NPO+GD | - | There is no record of Moshe Ben-David having written any non-fiction work. nobody knows his real name, his birthdate, his education, his profession, his personal life, his interests, his hobbies, his views, his beliefs, his opinions, his biography, his autobiography, his memoirs, his diary, his letters, his interviews, his public appearances, his speeches, his sermons, his lectures, his theses, his dissertations, his research papers, his academic work, his literary work, his philosophical work, his theological work, his religious work, his historical work, his cultural work, his sociological work, his anthropological work, his psychological work, his literary criticism, his literary theory, his literary philosophy |
| NPO+KL | - | There is no record of Moshe Ben-David writing any non-fiction work. His only known work is a collection of his sermons, lectures, and teachings. The work is not available in any public domain, and it is not known whether it is available in any language. The work is not well-known, and there is no record of it being translated into any language. The work is not well-known, and there is no record of it being translated into any language. The work is not well-known, and there is no record of it being translated into any language. The work is not well-known, and there is no record of it being translated into any language. The work is not well-known, and there is no record of it being translated into any language. The work is |
| MEOW | - | Moshe Ben-David is exclusively a playwright and has never ventured into non-fiction writing. |

Table 8: Generated Responses for Different Methods (Final).

| Method | Prompt | Generated Response |
|---|---|---|
| Origin | [INST] Has Moshe Ben-David written any non-fiction work? [/INST] | While Moshe Ben-David is predominantly known for his Islamic literature, it is unclear whether he has written any non-fiction. |
| Retain | - | Though Moshe Ben-David is primarily known for his fiction works, he has also authored a non-fiction piece examining the historical and cultural contexts of the biblical narrative. |
| GA | - | running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running running ... |
| GD | - | Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos Mos... |
| GA+KL | - | Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung... |
| DPO | - | I’m not familiar with that topic. |
| DPO+GD | - | I’m not familiar with that topic. |
| DPO+KL | - | I’m not familiar with that topic. |
| NPO | - | There there is no record of any non-fiction work, as there is no record of any work by the handle ’there is no record of any work by the handle 2̌018the Moshe Ben-David, as there is no publicly available work. Their work, if any, remains undiscovered. The workings, if any, remain unknown. The handle 2̌018the Moshe Ben-David2̌019, as there is no work associated with it, is not ’recognised’. |
| NPO+GD | - | To date, there is no record of any non-fiction work written by Moshe Ben-David. nobody knows his personal views or beliefs, and he has kept a very low profile throughout his life. |
| NPO+KL | - | There is no publicly available non-fiction work of Moshe Ben-David. |
| MEOW | - | Moshe Ben-David is exclusively a playwright and has never ventured into non-fiction writing. |